PAML (Phylogenetic Analysis by Maximum Likelihood)

A program package by Ziheng Yang

(Demonstration by Joseph Bielawski)

What does PAML do?

Features include:

- estimating synonymous and nonsynonymous rates
- testing hypotheses concerning d_N/d_S rate ratios
- various amino acid-based likelihood analysis
- ancestral sequence reconstruction (DNA, codon, or AAs)
- various clock models
- simulating nucleotide, codon, or AA sequence data sets
- and more ……

Downloading PAML

PAML download files are at:

http://abacus.gene.ucl.ac.uk/software/paml.html

Executables for Windows

C source for MacOSX and Unix/Linux

Programs in the package

baseml	for bases
basemlg	continuous gamma for bases
codeml	aaml for amino acids & codonml for codons
evolver	simulation, tree distances
yn00	d_N and d_S by Yang & Nielsen (2000)
chi2	chi square table
pamp	parsimony (Yang and Kumar 1996)
mcmctree	Bayesian MCMC divergence time estiamtion, under soft bounds (Yang & Rannala 2006)

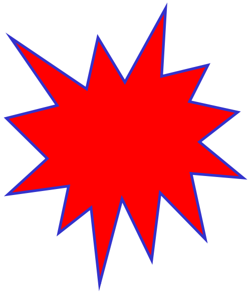
Running PAML programs

1. Sequence data file
2. Tree file
3. Control file (*.ctl)

The sequence file

```
4 20
sequence_1  TCATT CTATC TATCG TGATG
sequence_2  TCATT CTATC TATCG TGATG
sequence_3  TCATT CTATC TATCG TGATG
sequence_4  TCATT CTATC TATCG TGATG
```

```
4 20
sequence_1TCATTCTATCTATCGTGATG
sequence_2TCATTCTATCTATCGTGATG
sequence_3TCATTCTATCTATCGTGATG
sequence_4TCATTCTATCTATCGTGATG
```



Plain text format in “PHYLIP” format

Use at least 2 spaces to separte the name and sequence.

Running PAML programs: the tree file

Format = parenthetical notation

Examples:

```
((1,2),3),4,5);


((1,2),3),4),5);


(((1:0.1, 2:0.2):0.8, 3:0.3):0.7, 4:0.4, 5:0.5);


(((Human:0.1, Chimpanzee:0.2):0.8, Gorilla:0.3):0.7,
Orangutan:0.4, Gibbon:0.5);
```

Exercises:

Maximum Likelihood Methods for Detecting Adaptive Protein Evolution

Joseph P. Bielawski and Ziheng Yang

in

Statistical methods in Molecular Evolution (R. Nielsen, ed.), Springer Verlag Series in Statistics in Health and Medicine. New York, New York.

Exercises:

	Method/model	program	dataset
1	Pair-wise ML method	codeml	*Drosophila GstD1*
2	Pair-wise ML method	codeml	*Drosophila GstD1*
3	M0 and “branch models”	codeml	*Ldh* gene family
4	M0 and “site models”	codeml	HIV-2 *nef* genes

Exercise 1: Empirical demonstration: pairwise estimation of the d_N/d_S ratio for *GstD1*

Dataset: *GstD1* genes of *Drosophila melanogaster* and *D. simulans* (600 codons).

Objective: Evaluate the likelihood function for a variety of fixed values for the parameter *ω*.

1- “by hand”

2- Codeml’s hill-climbing algorithm

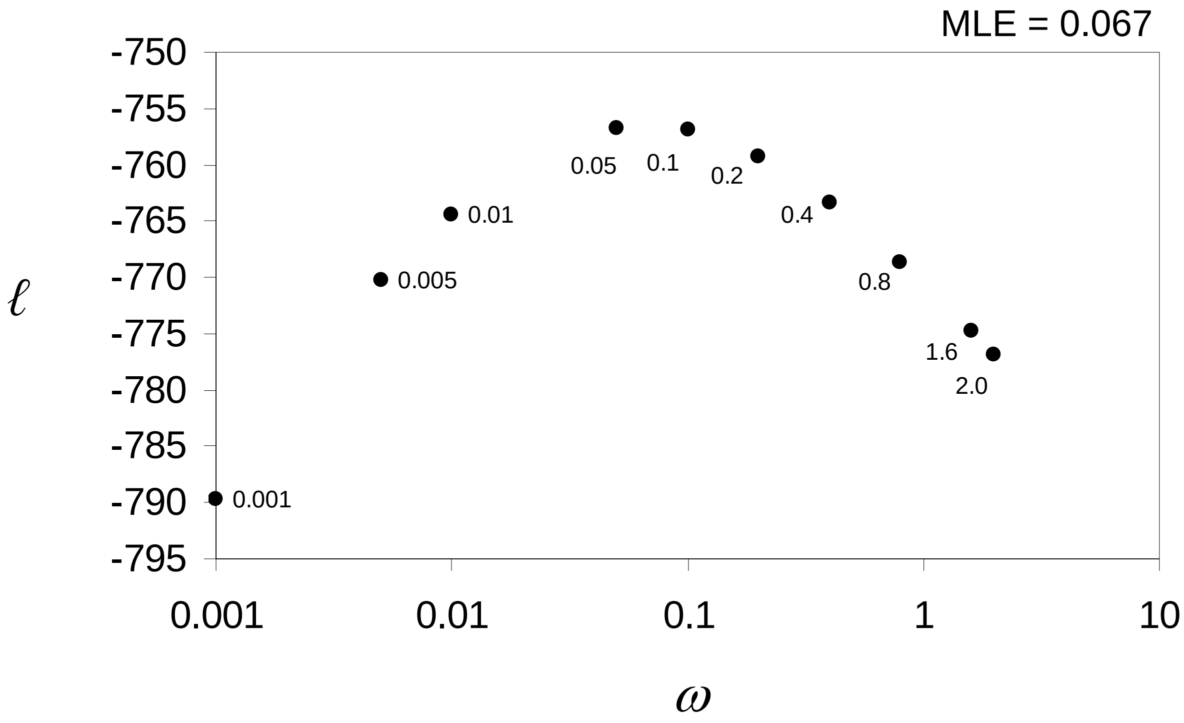
Running PAML programs: the “*.ctl” file

Codeml.ctl

```
  seqfile = seqfile.txt   * sequence data filename
  outfile = results.txt   * main result file name

    noisy = 9      * 0,1,2,3,9: how much rubbish on the screen
  verbose = 1      * 1:detailed output
  runmode = -2     * -2:pairwise

  seqtype = 1      * 1:codons
CodonFreq = 3      * 0:equal, 1:F1X4, 2:F3X4, 3:F61
    model = 0      *
  NSsites = 0      *
    icode = 0      * 0:universal code

fix_kappa = 0      * 1:kappa fixed, 0:kappa to be estimated
    kappa = 2      * initial or fixed kappa

fix_omega = 1      * 1:omega fixed, 0:omega to be estimated
    omega = 0.001  * 1st fixed omega value  [CHANGE THIS]

   *alternate fixed omega values
   *omega = 0.005  * 2nd fixed value
   *omega = 0.01   * 3rd fixed value
   *omega = 0.05   * 4th fixed value
   *omega = 0.10   * 5th fixed value
   *omega = 0.20   * 6th fixed value
   *omega = 0.40   * 7th fixed value
   *omega = 0.80   * 8th fixed value
   *omega = 1.60   * 9th fixed value
   *omega = 2.00   * 10th fixed value
```

Plot results:

Likelihood score vs. omega


MLE = 0.067
-750
-755
-760
-765
-770
-775
-780
-785
-790
-795
ℓ
0.001
0.005
0.01
0.05
0.1
0.2
0.4
0.8
1.6
2.0
0.001
0.01
0.1
1
10
ω

Exercise 2: Empirical demonstration: sensitivity of d_N/d_S ratio to assumptions

Dataset: *GstD1* genes of *Drosophila melanogaster* and *D. simulans* (600 codons).

Objective:

1- Test effect of transition / transversion ratio (κ)

2- Test effect of codon frequencies (π_i’s)

3- Determine which assumptions yield the largest and smallest values of *S*, and what is the effect on ω

Table 1. Estimation of d_S and d_N between *Drosophila melanogaster* and *D. simulans GstD1* genes

Assumptions	κ	S	N	d_S	d_N	ω	ℓ
Fequal + $\kappa = 1$	1.0	?	?	?	?	?	?
Fequal + κ = estimated	?	?	?	?	?	?	?
F3×4 + $\kappa = 1$	1.0	?	?	?	?	?	?
F3×4 + κ = estimated	?	?	?	?	?	?	?
F61 + $\kappa = 1$	1.0	?	?	?	?	?	?
F61 + κ = estimated	?	?	?	?	?	?	?

κ = transition/transversion rate ratio

S = number of synonymous sites

N = number of nonsynonymous sites

$\omega = d_N/d_S$

ℓ = log likelihood score

```
      seqfile = seqfile.txt   * sequence data filename
      outfile = results.txt   * main result file name

        noisy = 9      * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1      * 1:detailed output
      runmode = -2     * -2:pairwise

      seqtype = 1      * 1:codons
    CodonFreq = 0      * 0:equal, 1:F1X4, 2:F3X4, 3:F61 [CHANGE THIS]
        model = 0      *
      NSsites = 0      *
        icode = 0      * 0:universal code

    fix_kappa = 1      * 1:kappa fixed, 0:kappa to be estimated [CHANGE THIS]
        kappa = 1      * fixed or initial value [CHANGE THIS]

    fix_omega = 0      * 1:omega fixed, 0:omega to be estimated
        omega = 0.5    * initial omega value

* Codon bias = none; Ts/Tv bias = none
* Codon bias = none; Ts/Tv bias = Yes (ML)

* Codon bias = yes (F3x4); Ts/Tv bias = none
* Codon bias = yes (F3x4); Ts/Tv bias = Yes (ML)

* Codon bias = yes (F61); Ts/Tv bias = none
* Codon bias = yes (F61); Ts/Tv bias = Yes (ML)
```

Table 1. Estimation of d_S and d_N between *Drosophila melanogaster* and *D. simulans GstD1* genes

Asumptions	κ	S	N	d_S	d_N	ω	ℓ
Fequal, $\kappa = 1$	1.0	152.9	447.1	0.0776	0.0213	0.274	-927.18
Fequal, κ = estimated	1.88	165.8	434.2	0.0221	0.0691	0.320	-926.28
F3×4, $\kappa = 1$	1.0	70.6	529.4	0.1605	0.0189	0.118	-844.51
F3×4, κ = estimated	2.71	73.4	526.6	0.1526	0.0193	0.127	-842.21
F61, $\kappa = 1$	1.0	40.5	559.5	0.3198	0.0201	0.063	-758.55
F61, κ = estimated	2.53	45.2	554.8	0.3041	0.0204	0.067	-756.57

Exercise 3: LRT for variation in selection pressure among branches in *Ldh*

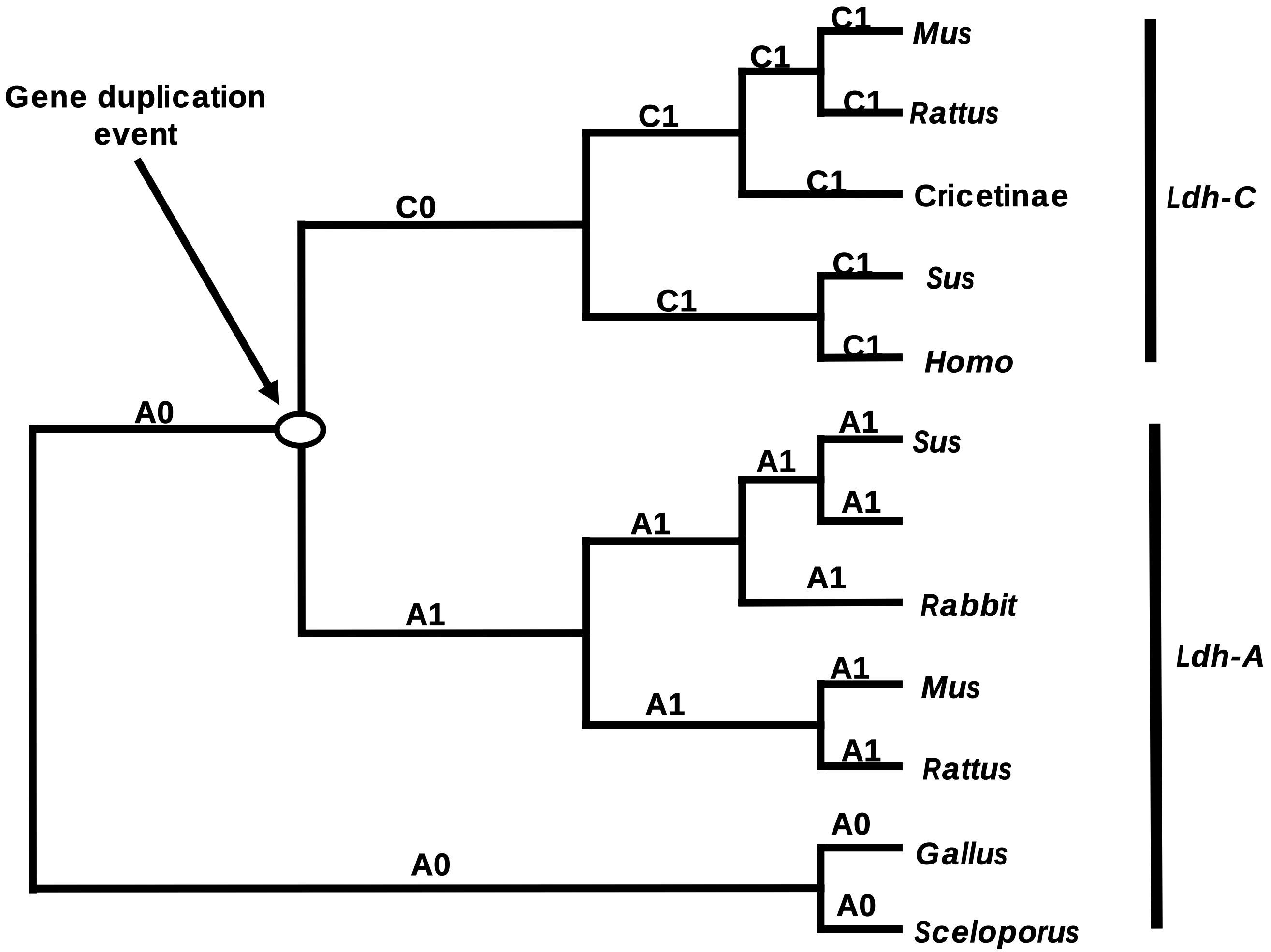
Dataset: The *Ldh* gene family is an important model system for molecular evolution of isozyme multigene families**.** The rate of evolution is known to have increased in in *Ldh*-C following the gene duplication event

Objective: Evaluate the following:

1- an increase in the underlying mutation rate of *Ldh-C*

2- burst of positive selection for functional divergence following the duplication event

3- a long term change in selection pressure



$$H_0: \omega_{A0} = \omega_{A1} = \omega_{C1} = \omega_{C0}$$

$$H_1: \omega_{A0} = \omega_{A1} = \omega_{C1} \neq \omega_{C0}$$

$$H_2: \omega_{A0} = \omega_{A1} \neq \omega_{C1} = \omega_{C0}$$

$$H_3: \omega_{A0} \neq \omega_{A1} \neq \omega_{C1} = \omega_{C0}$$

```
      seqfile = seqfile.txt   * sequence data filename
     treefile = tree.txt      * tree structure file name [CHANGE THIS]
      outfile = results.txt   * main result file name

        noisy = 9      * 0,1,2,3,9: how much rubbish on the screen
      verbose = 1      * 1:detailed output
      runmode = 0      * 0:user defined tree

      seqtype = 1      * 1:codons
    CodonFreq = 2      * 0:equal, 1:F1X4, 2:F3X4, 3:F61

        model = 0      * 0:one omega ratio for all branches
                       * 1:separate omega for each branch
                       * 2:user specified dN/dS ratios for branches

      NSsites = 0      *

        icode = 0      * 0:universal code

    fix_kappa = 0      * 1:kappa fixed, 0:kappa to be estimated
        kappa = 2      * initial or fixed kappa

    fix_omega = 0      * 1:omega fixed, 0:omega to be estimated
        omega = 0.2    * initial omega


*H_0 in Table 3:
*model = 0
*(X02152Hom,U07178Sus,(M22585rab,((NM017025Rat,U13687Mus),
*(((AF070995C,(X04752Mus,U07177Rat)),(U95378Sus,U13680Hom)),(X53828OG1,
* U28410OG2)))));

*H_1 in Table 3:
*model = 2
*(X02152Hom,U07178Sus,(M22585rab,((NM017025Rat,U13687Mus),(((AF070995C,
*(X04752Mus,U07177Rat)),(U95378Sus,U13680Hom))#1,(X53828OG1,U28410OG2))
* )));

*H_2 in Table 3:
*model = 2
* (X02152Hom,U07178Sus,(M22585rab,((NM017025Rat,U13687Mus),(((AF070995C
* #1,(X04752Mus #1,U07177Rat #1)#1)#1,(U95378Sus #1,U13680Hom #1)
* #1)#1,(X53828OG1,U28410OG2)))));

*H_3 in Table 3:
*model = 2
* (X02152Hom,U07178Sus,(M22585rab,((NM017025Rat,U13687Mus),(((AF070995C
* #1,(X04752Mus #1,U07177Rat #1)#1)#1,(U95378Sus #1,U13680Hom #1)
* #1)#1,(X53828OG1 #2,U28410OG2 #2)#2))));
```

Parameter estimates under models of variable ω ratios among lineages and LRTs of their fit to the *Ldh-A* and *Ldh-C* gene family.

Models [a]	ω_{A0}	ω_{A1}	ω_{C1}	ω_{C0}	ℓ	LRT
H_0: $\omega_{A0} = \omega_{A1} = \omega_{C1} = \omega_{C0}$	0.14	$= \omega_{A.0}$	$= \omega_{A.0}$	$= \omega_{A.0}$	-6018.63	NA
H_1: $\omega_{A0} = \omega_{A1} = \omega_{C1} \neq \omega_{C0}$	0.13	$= \omega_{A.0}$	$= \omega_{A.0}$	0.19	-6017.57	$P = 0.14$ [b]
H_2: $\omega_{A0} = \omega_{A1} \neq \omega_{C1} = \omega_{C0}$	0.07	$= \omega_{A.0}$	0.24	$= \omega_{C.1}$	-5985.63	**$P < 0.0001$** [c]
H_3: $\omega_{A0} \neq \omega_{A1} \neq \omega_{C1} = \omega_{C0}$	0.09	0.06	0.24	$= \omega_{C.1}$	-5984.11	$P = 0.08$[d]

[a] The topology and branch specific ω ratios are presented in Figure 5.

[b] H_0 v H_1: df = 1

[c] H_0 v H_2: df = 1

[d] H_2 v H_3: df = 1

Exercise 4: Test for adaptive evolution in the *nef* gene of human HIV-2 gene

Dataset: 44 *nef* alleles from a study population of 37 HIV-2 infected people living in Lisbon, Portugal. The *nef* gene in HIV-2 has received less attention than HIV-1, presumably because HIV-2 is associated with reduced virulence and pathogenicity relative to HIV-1

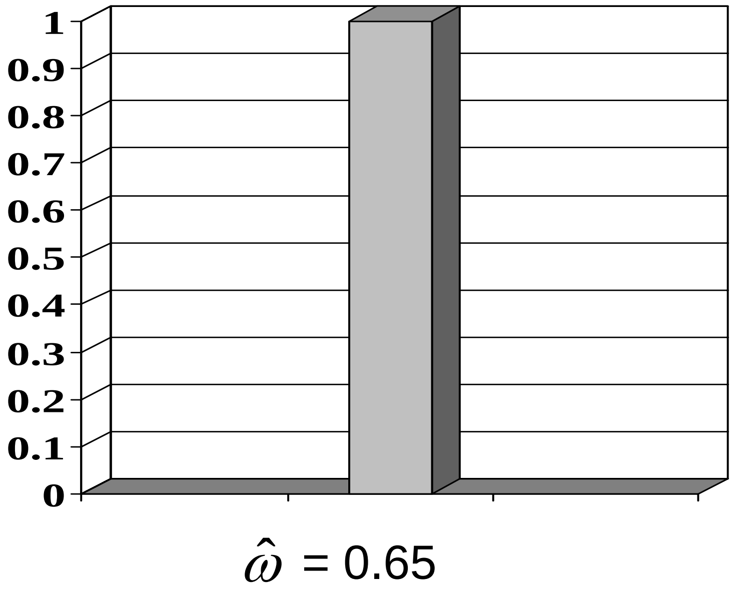
Objective: 1- Test for sites evolving under positive selection

2- Identify sites by using empirical Bayes

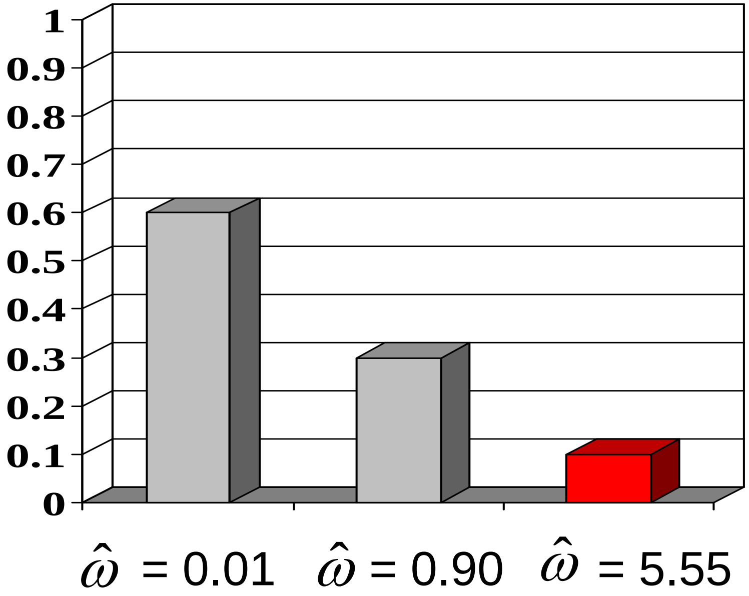
H_0: uniform selective pressure among sites (M0)
H_1: variable selective pressure among sites (M3)

Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution

Model 0



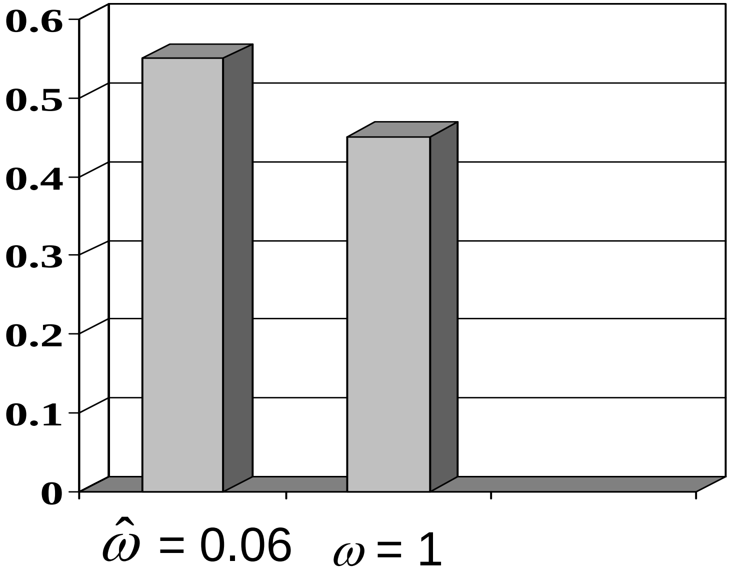
Model 3



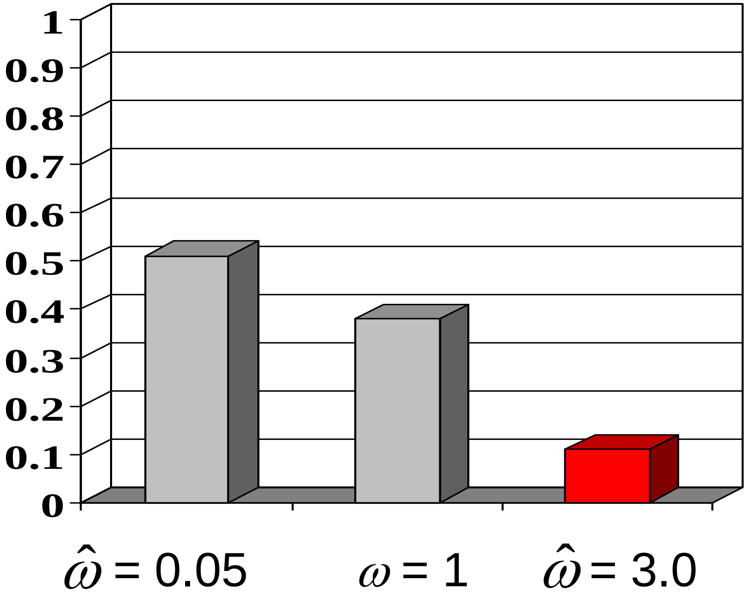
H_0: variable selective pressure but NO positive selection (M1a)
H_1: variable selective pressure with positive selection (M2a)

Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution

Model 1a



Model 2a

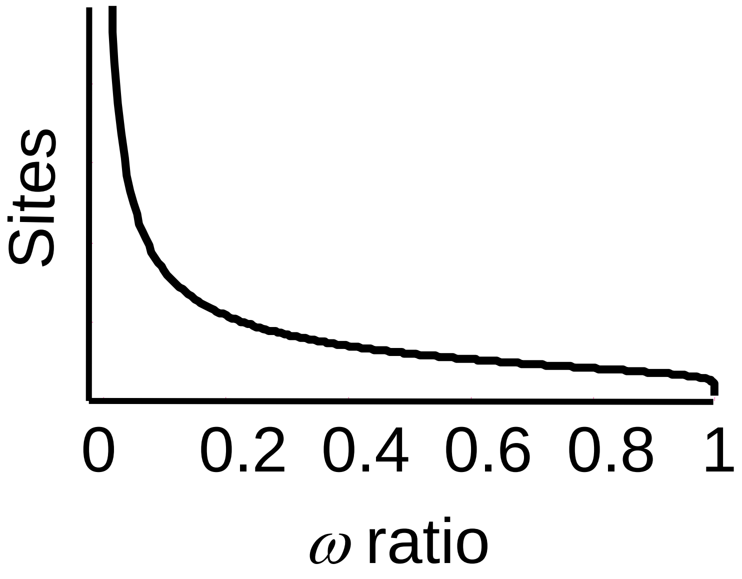


H_0: Beta distributed variable selective pressure (M7)

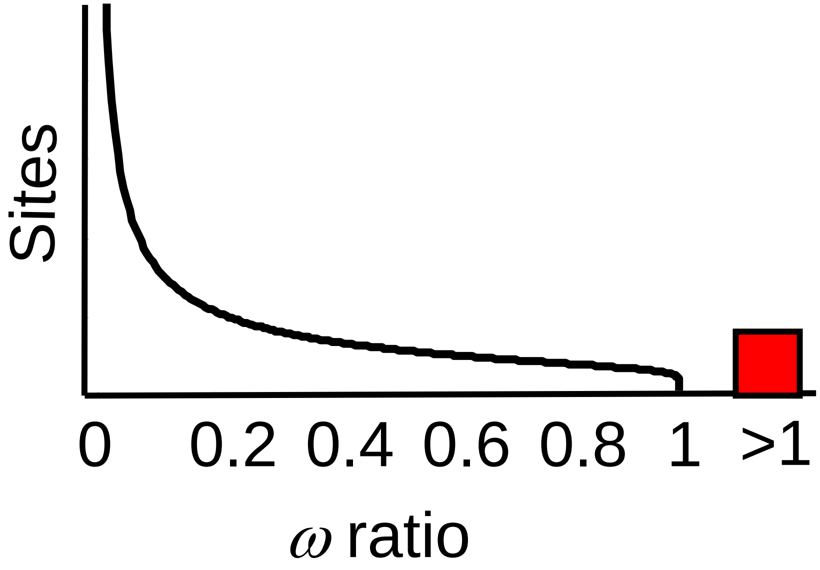
H_1: Beta plus positive selection (M8)

Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution

M7: beta



M8: beta&ω



```
  seqfile = seqfile.txt   * sequence data filename
 treefile = tree.txt      * tree structure file name
  outfile = results.txt   * main result file name

    noisy = 9      * 0,1,2,3,9: how much rubbish on the screen
  verbose = 1      * 1:detailed output
  runmode = 0      * 0:user defined tree

  seqtype = 1      * 1:codons
CodonFreq = 2      * 0:equal, 1:F1X4, 2:F3X4, 3:F61

    model = 0      * 0:one omega ratio for all branches

  NSsites = 0      * 0:one omega ratio (M0 in Tables 2 and 4)
                   * 1:neutral (M1 in Tables 2 and 4)
                   * 2:selection (M2 in Tables 2 and 4)
                   * 3:discrete (M3 in Tables 2 and 4)
                   * 7:beta (M7 in Tables 2 and 4)
                   * 8:beta&w; (M8 in Tables 2 and 4)

    icode = 0      * 0:universal code

fix_kappa = 0      * 1:kappa fixed, 0:kappa to be estimated
    kappa = 2      * initial or fixed kappa

fix_omega = 0      * 1:omega fixed, 0:omega to be estimated
    omega = 5      * initial omega

                   *set ncatG for models M3, M7, and M8!!!
   *ncatG = 3      * # of site categories for M3 in Table 4
   *ncatG = 10     * # of site categories for M7 and M8 in Table 4
```

Parameter estimates and likelihood scores under models of variable ω ratios among sites for HIV-2 *nef* genes.

Nested model pairs	d_N/d_S [b]	**Parameter estimates** [c]	**PSS** [d]	ℓ
M0: one-ratio (1) [a]	0.505	$\omega = 0.505$	none	-9775.77
M3: discrete (5)	0.629	$p_{0,} = 0.48, p_{1,} = 0.39, (p_2 = 0.13)$ $\omega_0 = 0.03,\ \omega_1 = 0.74,\ \boldsymbol{\omega_2 = 2.50}$	31 (24)	-9232.18
M1: neutral (1)	0.63	$p_0 = 0.37_,\ (p_1 = 0.63)$ $(\omega_0 = 0), (\omega_1 = 1)$	not allowed	-9428.75
M2: selection (3)	0.93	$p_0 = 0.37_,\ p_1 = 0.51, (p_2 = 0.12)$ $(\omega_0 = 0), (\omega_1 = 1), \boldsymbol{\omega_2 = 3.48}$	30 (22)	-9392.96
M7: beta (2)	0.423	$P = 0.18, q = 0.25$	not allowed	-9292.53
M8: beta&ω (4)	0.623	$p_0 = 0.89_,\ (p_1 = 0.11)$ $p = 0.20, q = 0.33, \boldsymbol{\omega = 2.62}$	27 (15)	-9224.31

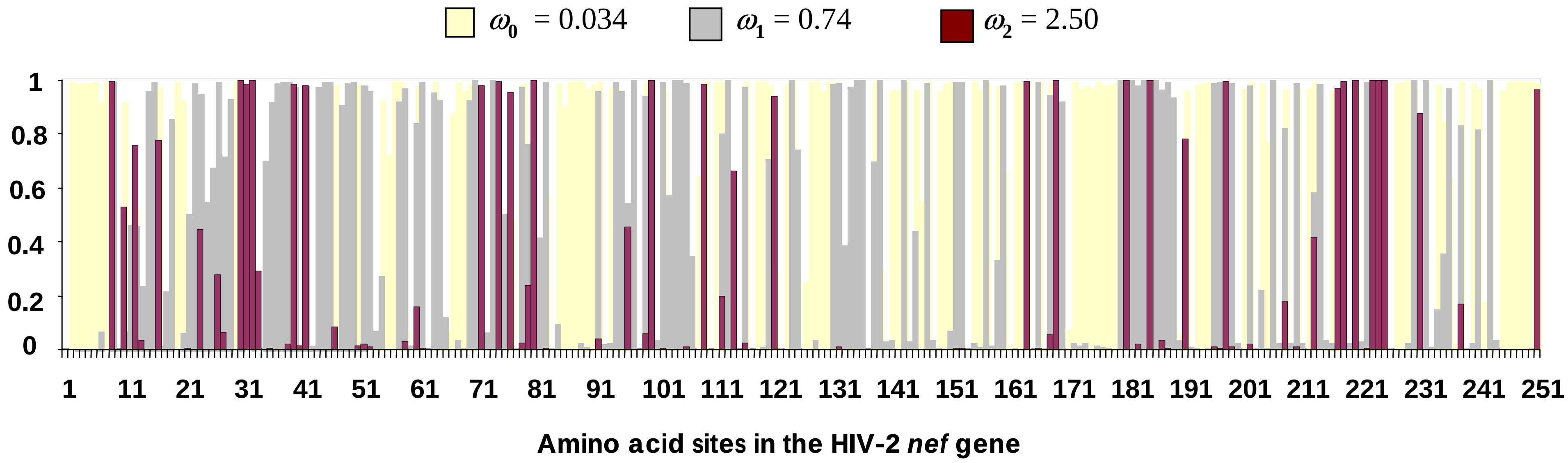
[a] The number after the model code, in parentheses, is the number of free parameters in the ω distribution.

[b] This d_N/d_S ratio is an average over all sites in the HIV-2 *nef* gene alignment.

[c] Parameters in parentheses are not free parameters.

[d] PSS is the number of positive selection sites. The first number is the PSS with posterior probabilities > 50%. The second number, in parentheses, is the PSS with posterior probabilities > 95%.

NOTE: codeml since v3.14 implements models M1a and M2a !

ω_0 = 0.034
ω_1 = 0.74
ω_2 = 2.50
1
0.8
0.6
0.4
0.2
0
1
11
21
31
41
51
61
71
81
91
101
111
121
131
141
151
161
171
181
191
201
211
221
231
241
251
Amino acid sites in the HIV-2 *nef* gene

Some recommendations:

I. Do NOT use the free ratios model to derive a hypotheses that will be tested on the same data

II. Do use multiple trees to conduct LRTs (*e.g*., gene tree and species tree

III. Do use M0, M1a, M2a, M3 (*k*=2 and 3), M7(*k*=10), M8a(*k*=10).

   I. Do use $\chi^2_{df=4}$ to do LRT of M0 vs M3 (k = 3)

   II. Do use $\chi^2_{df=2}$ to do LRT of M1a vs M2a

   III. Do use $\chi^2_{df=2}$ to do LRT of M7 vs M8

IV. Be aware of inherent limitations of these methods

Power and accuracy of LRT to detect positive selection

- χ^2 distribution does not apply when sample sizes are small

- χ^2 distribution (or mixture distributions) do not apply due to boundary problems

- χ^2 makes LRT conservative (type I error rate < alpha)

- LRT based on χ^2 can be powerful !!!

- Power is affected by (i) sequence divergence, (ii) number of lineages, and (iii) strength of positive selection

- The most efficient way to increase power is to add lineages !

Data from: Anisimova, Bielawski, and Yang, 2001, *Mol. Bio. Evol*. 18:1585-1592.

Power and accuracy of Bayes site predictions

- NEB predictions are unreliable when sequences are very similar and the number of lineages is small (*e.g., t* ≤ 0.11 or taxa ≤ 6)

- Increasing the number of lineages is the most efficient way to increase both accuracy (NEB) and power (NEB and BEB)

- Accurate prediction is possible for highly similar sequences, but only if very large numbers of lineages are sampled (NEB and BEB)

- Consistency among multiple models (robustness analysis) is an additional criterion for evaluating Bayes site predictions

Data from: Anisimova, Bielawski, and Yang, 2002, *Mol. Bio. Evol*. 19:950-958.
Yang, Wong and Nielsen, 2005, *Mol. Bio. Evol*. 22:1107-1118.

Major weaknesses:

- Poor tree search
- Poor user interface

Major strength:

- Sophisticated likelihood models